Marker-Assisted Backcrossing for Introgression of a Recessive Gene

Matthias Frisch and Albrecht E. Melchinger*

ABSTRACT

Molecular markers are used to trace the presence of target genes (foreground selection) and accelerate recovery of the recurrent parent genome (background selection) in backcross programs. In this study, we present an approach for introgression of a recessive target gene from a donor into the genetic background of a recipient line by foreground selection combined with background selection for reducing the donor chromosome segment around the target gene. The goal of the proposed breeding plan is to generate with a given probability, q_2, up to the second backcross generation (BC_2) at least $k \geq 1$ individuals, which carry the target gene and are homozygous for the recurrent parent alleles at flanking markers, by means of a minimum number of individuals. We provide formulas for calculation of (i) the population size required in generation BC_1 and (ii) the probability of success q_2 of the breeding program in generation BC_2. The latter depends on the number and genotype of the BC_1 individuals selected for further backcrossing and the size of their BC_2 families. The optimum allocation of individuals to generations BC_1 and BC_2 was determined by computer simulations for various map distances between the target gene and the flanking markers. Our approach is demonstrated by a numerical example and can assist breeders in the optimum design of breeding programs for marker-assisted introgression of a recessive gene.

Many important genes in breeding for resistance and quality traits are inherited recessively. In conventional backcross programs for introgression of a recessive target gene, that gene's presence or absence in a backcross individual is determined by a phenotypic assay of progeny generated either by selfing or by crossing to the donor parent (Allard, 1960). As an alternative to this time-consuming method, flanking molecular markers can be used as a diagnostic tool to trace the presence of the target gene (foreground selection) in successive backcross generations. By this approach, presence of the target gene must be tested either by selfing or crossing to the donor only at the end of the breeding program. In addition, markers with a good coverage of the entire genome can be used to select for rapid recovery of the recurrent parent genome (background selection).

Marker-assisted foreground selection was proposed by Tanksley (1983) and investigated in the context of introgression of resistance genes by Melchinger (1990), who presented an a priori approach for calculating the minimum number of individuals and family size required in recurrent backcrossing. Marker-assisted background selection was proposed by Young and Tanksley (1989) and investigated by various authors (Hospital et al., 1992; Openshaw et al., 1994; Visscher et al., 1996; Frisch et al., 1999a,b). Hospital and Charcosset (1997) investigated combined foreground and background selection for introgression of favorable genes at quantitative trait loci (QTL). They presented an a priori approach to calculate the required population size for the case of several target loci and map positions estimated with varying accuracy. To our knowledge, no previous study exists concerning combined foreground and background selection for introgression of a recessive gene with known map position.

The objectives of this study were to (i) devise a breeding plan for combined foreground and background selection for introgression of a recessive gene, (ii) provide formulas for calculation of the required population size in generation BC_1, (iii) derive the probability of success of the breeding program in generation BC_2 depending on the number and genotype of the BC_1 individuals selected for further backcrossing and family size of their BC_2 progeny, and (iv) present simulation results with respect to the optimum allocation of resources in generations BC_1 and BC_2 for various distances between the target gene and flanking markers. Following Frisch et al. (1999a), we adopted an a posteriori approach in which the design of generation BC_2 is determined on the basis of the known marker genotypes of the BC_1 individuals selected for further backcrossing.

MODEL

Assumptions

Under the assumptions (a) the average number of crossovers formed on a chromatid is equal to its length in Morgan units and (b) the locations of crossovers are uniformly and independently distributed on the chromatid, the number of crossovers formed in a chromatid segment of length d follows a Poisson distribution with parameter d. Assumptions (a) and (b) imply that neither chiasma interference nor chromatid interference occur (Stam, 1979). Furthermore, the probability p_r of recombination between two loci is related to their map distance d (in Morgan units) by Haldane's (1919) mapping function

$$p_r = (1 - e^{-2d})/2. \qquad [1]$$

Definitions

We consider a chromosome of length L. Positions on the chromosome are represented by a scale (in Morgan units) ranging from 0 to L. The target locus is located at position x and two flanking markers, used for foreground selection, are located at positions m_l and m_r (Fig. 1). If only one marker is used for foreground selection, we assume without loss of generality that it is located at position m_r. Two markers located at positions y_l and y_r are used for background selection. Let $d_1 = x - y_l$, $d_2 = y_r - x$, $\delta_1 = x - m_l$, $\delta_2 = m_r - x$ with $\delta_1 < d_1$ and $\delta_2 c d_2$. The events A to H refer to the occurrence of recombination (i.e., an odd number of crossovers) between

Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70953 Stuttgart, Germany. *Corresponding author (melchinger@uni-hohenheim.de).

Received 31 July 2000.

Published in Crop Sci. 41:1485–1494 (2001).

Abbreviations: BC_t, t-th backcross generation; cM, centimorgan; M, Morgan; QTL, quantitative trait loci.

Table 1. Formulas to calculate the probabilities $p_{0,g}$ (probability that a BC$_1$ individual has marker genotype g), $p_{g+|0,g}$ (probability that a BC$_1$ individual with marker genotype g carries the target gene), and $P_{g+,T+}$ (probability that a BC$_1$ individual with marker genotype g^+ generates a BC$_2$ individual with marker genotype $t^+ \in T^+$); see text for detailed definitions of $p_{0,g}$, $p_{g+|0,g}$, and $p_{g+,T+}$.

Marker genotype $g \in G$	$p_{0,g}$	$p_{g+\|0,g}$	$p_{g+,T+}$
		One foreground selection marker	
$y_l^- m_r^+ y_r^-$†	$p_g p_e/2$‡	$p_a(1 - p_d)/p_g$	$(1 - p_d)/2$
$y_l^+ m_r^+ y_r^-$	$(1 - p_g)p_e/2$	$(1 - p_a)(1 - p_d)/(1 - p_g)$	$p_a(1 - p_d)/2$
$y_l^- m_r^+ y_r^+$	$p_g(1 - p_e)/2$	$p_a(1 - p_d)/p_g$	$(1 - p_d)p_e/2$
		Two foreground selection markers	
$y_l^- m_l^+ m_r^+ y_r^-$	$p_b(1 - p_h)p_e/2$	$(1 - p_c)(1 - p_d)/(1 - p_h)$	$(1 - p_c p_d)/2$
$y_l^- m_l^- m_r^+ y_r^-$	$(1 - p_b)p_h p_e/2$	$p_c(1 - p_d)/p_h$	$(1 - p_d)/2$
$y_l^- m_l^+ m_r^- y_r^-$	$p_b p_h(1 - p_e)/2$	$(1 - p_c)p_d/p_h$	$(1 - p_c)/2$
$y_l^+ m_l^+ m_r^+ y_r^-$	$(1 - p_b)(1 - p_h)p_e/2$	$(1 - p_c)(1 - p_d)/(1 - p_h)$	$[p_b(1 - p_c) + (1 - p_b)p_c(1 - p_d)]/2$
$y_l^- m_l^+ m_r^+ y_r^+$	$p_b(1 - p_h)(1 - p_e)/2$	$(1 - p_c)(1 - p_d)/(1 - p_h)$	$[(1 - p_d)p_e + (1 - p_c)p_d(1 - p_e)]/2$
$y_l^+ m_l^- m_r^+ y_r^-$	$p_b p_h p_e/2$	$p_c(1 - p_d)/p_h$	$p_a(1 - p_d)/2$
$y_l^+ m_l^+ m_r^- y_r^-$	$(1 - p_b)p_h(1 - p_e)/2$	$(1 - p_c)p_d/p_h$	$p_b(1 - p_c)/2$
$y_l^- m_l^- m_r^+ y_r^+$	$(1 - p_b)p_h(1 - p_e)/2$	$p_c(1 - p_d)/p_h$	$(1 - p_d)p_a/2$
$y_l^- m_l^+ m_r^- y_r^+$	$p_b p_h p_e/2$	$(1 - p_c)p_d/p_h$	$(1 - p_c)p_f/2$

† The symbols y_l and y_r denote the background selection makers, m_l and m_r the foreground selection markers, and x the target locus. A superscript + or − indicates that the locus is heterozygous or homozygous for the recurrent parent allele, respectively.

‡ See text for definitions of probabilities p_a to p_h.

the loci delimiting the intervals $[y_l, x]$, $[y_l, m_l]$, $[m_l, x]$, $[x, m_r]$, $[m_r, y_r]$, $[x, y_r]$, $[y_l, m_r]$, and $[m_l, m_r]$, respectively. The corresponding probabilities p_a to p_h can be obtained from Eq. [1] by inserting the map distance between the loci delimiting the respective interval.

Adopting the termimology of Hospital and Charcosset (1997), we denote by c^- the genotype of an individual homozygous for the recurrent parent allele and by c^+ the genotype of an individual heterozygous for the recurrent parent allele at the locus at position c ($c \in \{y_l, m_l, x, m_r, y_r\}$). We further define indicator variables Y_l, M_l, X, M_r, and Y_r, which take the value 1 if the marker at the respective position is heterozygous and 0 if it is homozygous for the recurrent parent allele.

Let A denote the set of markers employed in a backcross program. We have for one foreground selection marker $A = \{y_l, m_r, y_r\}$ and for two foreground selection markers $A = \{y_l, m_l, m_r, y_r\}$. The set Ω_A contains all possible multilocus marker genotypes for a set of markers ($|\Omega_A| = 2^3 = 8$ for one foreground selection marker and $|\Omega_A| = 2^4 = 16$ for two foreground selection markers).

By definition, the subset $G \subseteq \Omega_A$ contains all multilocus marker genotypes with at least one background selection marker homozygous for the recurrent parent allele and at least one heterozygous foreground selection marker

$$G = \{\omega_A \in \Omega_A \mid Y_l + Y_r \leq 1 \wedge M_l + M_r \geq 1\}. \qquad [2]$$

The elements of G are listed in Table 1. In addition, we define the Genotype 0 consisting of heterozygous F$_1$ individuals and $G_0 = G \cup \{0\}$.

The subset $T \subseteq G$ contains all multilocus marker genotypes with both background selection markers homozygous for the recurrent parent allele and at least one heterozygous foreground selection marker

$$T = \{\omega_A \in \Omega_A \mid Y_l + Y_r = 0 \wedge M_l + M_r \geq 1\}. \qquad [3]$$

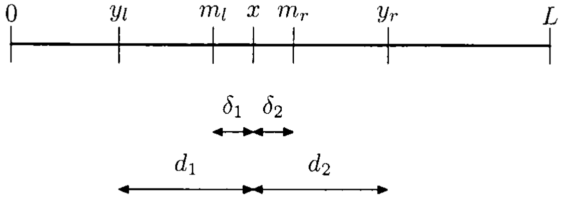

Fig. 1. Chromosome of length L with the target locus at position x, two markers for foreground selection at positions m_l and m_r, and two markers for background selection at positions y_l and y_r. The map distances of the target locus to the foreground selection markers are denoted with δ_1 and δ_2 and those to the background selection markers with d_1 and d_2.

Let $B = A \cup \{x\}$ and Ω_B the set of all possible multilocus genotypes with respect to B. Thus, $|\Omega_B| = 16$ for one foreground selection marker and $|\Omega_B| = 32$ for two foreground selection markers. By the same token as above, we define the following two subsets for carriers of the target gene:

$$G^+ = \{\omega_B \in \Omega_B \mid Y_l + Y_r \leq 1 \wedge M_l + M_r \geq 1 \wedge X = 1\} \quad [4]$$

$$T^+ = \{\omega_B \in \Omega_B \mid Y_l + Y_r = 0 \wedge M_l + M_r \geq 1 \wedge X = 1\} \quad [5]$$

Elements of the sets G, T, G^+, and T^+ are denoted with the lowercase letters g, t, g^+, and t^+, respectively.

We define the following probabilities.

1. $p_{0,g}$: Probability that a BC$_1$ individual has marker genotype g.
2. $p_{0,g+}$: Probability that a BC$_1$ individual has marker genotype g^+.
3. $p_{g+|0,g}$: Conditional probability that a BC$_1$ individual carries the target gene under the condition that it has marker genotype g.
4. $p_{g,T+}$: Probability that a backcross progeny of an individual with genotype g has a genotype $t^+ \in T^+$.
5. $p_{g+,T+}$: Probability that a backcross progeny of an individual with genotype g^+ has a genotype $t^+ \in T^+$.

Equations for calculating the probabilities $p_{0,g}$, $p_{g+|0,g}$, and $p_{g+,T+}$ from p_a to p_h are given in Table 1. The probabilities $p_{0,g+}$ and $p_{g,T+}$ can be calculated as

$$p_{0,g+} = p_{0,g}\, p_{g+|0,g} \qquad [6]$$

$$P_{g,T+} = p_{g+|0,g}\, p_{g+,T+}. \qquad [7]$$

For further derivations, we also need the probabilities $p_{0,T+}$ that a F$_1$ individual generates by backcrossing an individual of marker genotype $t^+ \in T^+$. For one foreground selection marker, we have

$$p_{0,T+} = p_a(1 - p_d)p_e/2 \qquad [8]$$

and for two foreground selection markers

$$
\begin{aligned}
p_{0,T+} &= p_b(1 - p_c)(1 - p_d)p_e/2 \\
&\quad + (1 - p_b)p_c(1 - p_d)p_e/2 \\
&\quad + p_b(1 - p_c)p_d(1 - p_e)/2
\end{aligned} \qquad [9]
$$

Basic Result on the Minimum Population Size

If a particular genotype occurs with probability p, the number m of individuals of this type in a sample of size n is binomially distributed with probability

$$B(n, m, p) = \binom{n}{m} p^m (1 - p)^{n-m}. \qquad [10]$$

The minimum population size n required to find with probability q at least one individual of a genotype, which occurs with probability p, can be derived from Eq. [10] as

$$n \geq \ln(1 - q)/\ln(1 - p). \qquad [11]$$

BREEDING PLAN

For introgression of a recessive gene with combined foreground and background selection for reduction of the intact donor chromosome segment around the target gene, we propose a breeding plan designed for producing at least one individual of genotype $t^+ \in T^+$ at latest in generation BC$_2$ with a minimum expenditure. Such a breeding plan is fully described by answering the following questions.

1. What is the necessary population size n_1 in BC$_1$?
2. Suppose the marker genotypes of the n_1 BC$_1$ individuals are known. Which marker genotypes and how many individuals of each should be selected as parents for further backcrossing?
3. What should be the size f_g of a BC$_2$ family produced from a selected BC$_1$ individual of genotype g?

Population Size and Marker Analyses in BC$_1$

Our approach for choosing n_1 rests upon the fact that even with a large population size of several hundred individuals, the chances of finding a BC$_1$ individual of genotype $t^+ \in T^+$ are small because this requires double crossovers in a small chromosome region. In most cases, the overall goal is reached by recombination between the target gene and a background selection marker on one side in generation BC$_1$ and an analogous recombination on the other side of the target gene in generation BC$_2$. Hence, a realistic goal for generation BC$_1$ is to produce at least one individual which is (i) heterozygous for at least one foreground selection marker, (ii) homozygous for at least one background selection marker, and (iii) a carrier of the target gene. These three conditions are fulfilled by any individual with multilocus genotype $g^+ \in G^+$, but only the first two conditions can be determined by marker assays.

The minimum sample size n_1 to assure with probability q_1 that at least one individual of genotypes $g^+ \in G^+$ occurs in the BC$_1$ population is derived from Eq. [11] as

$$n_1 \geq \frac{\ln(1 - q_1)}{\ln\left(1 - \sum\limits_{g^+ \in G^+} p_{0,g+}\right)}. \qquad [12]$$

A generalization of Eq. [12], which allows one to determine n_1 to assure with probability q_1 the presence of k individuals of genotype $g^+ \in G^+$, is presented in Appendix A.

The BC$_1$ individuals are first analyzed for presence of the donor allele at the foreground selection marker(s). Individuals carrying the donor allele for at least one foreground selection marker are analyzed subsequently for the background selection markers. All BC$_1$ individuals with marker genotype $g \in G$ are potential parents for generation BC$_2$.

Selection of BC$_1$ Individuals and Family Size in Generation BC$_2$

When several individuals with different genotype $g \in G$ are found in the BC$_1$ population, the experimenter must decide which and how many of them should be used as parents for producing the BC$_2$ generation. This choice should be subject to the condition that a desired probability of success q_2 is reached with a minimum number of individuals. Let denote (i) $(o_g)_{g \in G}$ the number of individuals with genotype g observed in BC$_1$, (ii) $(i_g)_{g \in G}$ the number of BC$_1$ individuals with genotype g used for further backcrossing, and (iii) $(f_g)_{g \in G}$ the size of a BC$_2$ family produced from a BC$_1$ individual with genotype g. If only few BC$_1$ individuals with genotype $g \in G$ are found and $p_{g+|0,g}$ or $p_{g+,T+}$ are small, it may be necessary to back up one generation and backcross F$_1$ individuals to the recurrent parent. We denote the respective parameters with i_0 and f_0 and set $o_0 = 1$.

A certain parameter setting for generating the BC$_2$ generation, consisting of the number individuals to be backcrossed i_g and the respective family size f_g for each marker genotype $g \in G_0$, is denoted by $(i_g, f_g)_{g \in G_0}$. The set $\mathscr{A}$ of all admissible parameter settings $(i_g, f_g)_{g \in G_0}$ is determined by $(o_g)_{g \in G}$, the maximum possible family size m (which can be determined either by the multiplication rate of the species of the resources of the breeder), and the desired probability of success q_2

$$
\begin{aligned}
\mathscr{A}((o_g)_{g \in G}, m, q_2) = \{&(i_g, f_g)_{g \in G_0} \mid 0 \leq i_g \leq o_g, \\
&0 \leq f_g \leq m, q((i_g, f_g)_{g \in G_0}) \geq q_2\}
\end{aligned} \qquad [13]
$$

The probability $q[(i_g, f_g)_{g \in G_0}]$ of recovering at least one progeny of genotype $t^+ \in T^+$ when using the parameter setting $(i_g, f_g)_{g \in G_0}$ is calculated as

$$q((i_g, f_g)_{g \in G_0}) = 1 - \prod_{g \in G_0} [1 - q_g(i_g, f_g)] \qquad [14]$$

where $q_g(i_g, f_g)$ is the probability of finding among the i_g BC families of size f_g at least one individual with genotype $t^+ \in T^+$

$$
\begin{aligned}
q_g(i_g, f_g) &= \sum_{s=1}^{i_g} \{B(i_g, s, p_{g+|0,g}) \\
&\qquad \times [1 - B(sf_g, 0, p_{g+,T+})]\} \\
&= 1 - B(i_g, 0, p_{g+|0,g}) \\
&\qquad - \sum_{s=1}^{i_g} B(i_g, s, p_{g+|0,g}) B(sf_g, 0, p_{g+,T+}).
\end{aligned} \qquad [15]
$$

In Appendix A, we give an extension of Eq. [14] which can be used for calculating the probability that at least k individuals of genotypes $t^+ \in T^+$ are found with the parameter setting $(i_g, f_g)_{g \in G_0}$.

The number of individuals required for the parameter setting $(i_g, f_g)_{g \in G_0}$ is

$$n_2((i_g, f_g)_{g \in G_0}) = \sum_{g \in G_0} i_g f_g \qquad [16]$$

and the optimum parameter setting $(i_g^*, f_g^*)_{g \in G_0}$ is the one requiring the smallest number of individuals among all elements in $\mathscr{A}$

$$n_2((i_g^*, f_g^*)_{g \in G_0}) = \min_{(i_g, f_g)_{g \in G_0} \in \mathcal{A}} n_2((i_g, f_g)_{g \in G_0}). \quad [17]$$

There is no closed analytical solution for the minimization problem in Eq. [17]. For finding a suitable parameter setting we propose to calculate the probability of success $q((i_g, f_g)_{g \in G_0})$ (Eq. [14]) for various parameter settings $(i_g, f_g)_{g \in G_0}$ and choose the one, which is element of $\mathcal{A}$ and requires the smallest number of individuals.

Before calculating $q((i_g, f_g)_{g \in G_0})$ for alternative parameter settings, it is useful to order the marker genotypes observed in the BC_1 population with respect to their probability of obtaining a backcross progeny of genotype $t^+ \in T^+$ as follows:

$$h \succ g \quad \text{if and only if} \quad p_{h,T^+} \geq p_{g,T^+} \quad [18]$$
$$\text{for } g, h \in G_0.$$

This provides a clue as to which marker genotypes should be preferably backcrossed. A more adequate ordering of the genotypes with respect to their contribution to the total probability of success (Eq. [14]) can be obtained when preliminary information about the number i_g of individuals to be backcrossed and the family size f_g to be used is available by defining

$$h \succ g \quad \text{if and only if} \quad q_h(i_h, f_h) \geq q_g(i_g, f_g),$$
$$\text{for } g, h \in G_0, \quad [19]$$

where $i_h f_h = i_g f_g$.

Marker Analysis of BC_2 Individuals and Progeny Testing

All BC_2 individuals are marker assayed at the markers heterozygous in their nonrecurrent BC_1 parent. BC_2 individuals with marker genotype $t \in T$ are either selfed or backcrossed to the donor genotype. Presence of the target gene in a backcross individual is determined by phenotypic evaluation of its progeny obtained wither by selfing or crossing with the donor parent.

DISCUSSION

Genetic Model

Following earlier studies (Hospital et al., 1992; Visscher et al., 1996; Hospital and Charcosset, 1997), we used Haldane's (1919) mapping function for modeling crossover formation during meiosis. It is well known that this is a simplified model because of the assumption of no interference (Stam, 1979). Since Haldane's pioneering paper, numerous researchers (e.g., Kosambi, 1944; Karlin and Liberman, 1978; Zhao and Speed, 1996; Browning, 2000) proposed alternative mathematical models which include interference. Most of the resulting map functions can be modeled by a stationary renewal process, the interevent distribution of which can be approximated by gamma distributions (Zhao and Speed, 1996). McPeek and Speed (1995) compared the fit of various crossover formation models and concluded that gamma interevent distribution fit best the *Drosophilia* dataset of Morgan et al. (1935).

We used Haldane's (1919) mapping function because of its mathematical simplicity and the stochastic independence of crossover formations in adjacent chromosome regions which allowed us to derive closed analytical formulas for the problems addressed in this paper. Applying gamma interevent distributions would in most instances yield unwieldy formulas which could only be numerically approximated. Moreover, as pointed out by Stam and Zeven (1981), dropping the assumption of no interference would reduce the generality of the presented approach because it would be necessary to know the type and degree of interference for the chromosome region of each target gene.

Under the assumption of positive chiasma interference (Stam, 1979), multiple crossovers in a given chromosome region occur less frequently than under the assumption of no interference. Consequently, if the target gene is located in a region with positive interference, the population sizes obtained by our equations are underestimated. The reverse holds true under the assumption of negative interference. In conclusion, the reader should be aware that the model presented (as with most mathematical models of biological systems) is not capable capturing every detail of the underlying biological process and the results presented should be interpreted with this in mind.

Comparison with Earlier Studies

Introgression of a recessive gene with combined foreground and background selection can be regarded as a special case of QTL introgression investigated by Hospital and Charcosset (1997) (one QTL with a zero-length confidence interval, two foreground selection markers, two background selection markers). Their Eq. [A.16] through [A.22] could be used to calculate the required population sizes of a breeding program for introgression of a recessive gene. Our approach differs from that of Hospital and Charcosset (1997) in three respects: (i) the definition of the goal of the breeding program, (ii) the selection strategy, and (iii) calculation of the population size required in each BC generation.

Concerning the goal of the breeding program, we propose to choose n_2 such that at least one BC_2 individual with genotype $t^+ \in T^+$ is obtained with probability q_2. In contrast, Hospital and Charcosset (1997) use in their Eq. [A.16] through [A.22] the probability of finding at least one individual with marker genotype $y_l^- m_l^+ m_r^+ y_r^-$, but they do not include a condition about presence of the target gene in their criterion. (Note: By modifying the definition of the probability P_M used in their paper, their approach could be used to determine n_1 and n_2 to generate with a certain probability at least one individual with genotype $t^+ \in T^+$.)

The main differences with respect to the selection strategy are (i) we propose to select as many promising BC_1 individuals as required to each a desired probability of success q_2 in generation BC_2, while Hospital and Charcosset (1997) based their calculations on selection of a single BC_1 individual; and (ii) we consider all BC_1 individuals with marker genotype $g \in G$ as potential parents for producing generation BC_2 and select individuals of one or several genotypes $g \in G_0$, on the basis of their effect on the probability $q((i_g, f_g)_{g \in G_0})$, i.e., depending on the marker distances d_1, d_2, δ_1, and δ_2. In contrast, Hospital and Charcosset (1997) propose to select an individual with all foreground selection markers carrying the donor allele and they do not distinguish between individuals of genotype y_l^- and y_r^-, even when $d_1 \neq d_2$.

In our approach, the number of BC_1 individuals selected for further backcrossing and the respective family size of their BC_2 progeny are determined after knowing the marker genotype of the BC_1 individuals (i.e., a posteriori). In contrast, Hospital and Charcosset (1997) propose to calculate the population size for all backcross generations before starting the breeding program (i.e., a priori). Taking into account the marker genotype of the selected BC_1 individual(s) has the following advantages: (i) only the number of BC_2 individuals actually required to ascertain a given probability of success q_2 is generated, and (ii) the desired probability of success q_2 is reached irrespective of the outcome in generation BC_1. Both properties follow directly from the Theorem of Total Probability.

The advantages of the a posteriori approach were previously demonstrated for the simpler case of marker-assisted background selection in combination with phenotypic selection for a dominant target gene (Frisch et al., 1999b). We give here only a short numerical example. Assume $d_1 = d_2 = 0.10$ M, $\delta_1 = 0.03$ M, and $\delta_2 = 0.05$ M. With our approach, the optimum population sizes to find with probability $q_2 = 0.99$ at least one BC_2 individual with genotype $t^+ \in T^+$ are $n_1 = 77$ and $n_2 = 102$ (Table 3). Applying the approach of Hospital and Charcosset (1997), the optimum population sizes to find with probability 0.99 at least one BC_2 individual of genotype $y_l^- m_l^+ m_r^+ y_r^-$ are $n_1 = 106$ and $n_2 = 188$. Besides requiring more than 100 additional individuals, an individual with marker genotype $y_l^- m_l^+ m_r^+ y_r^-$ carries only with probability $(1 - p_c)(1 - p_d)/(1 - p_h) = 0.88$ the target gene.

Direct selection for presence of a dominant target gene combined with marker-assisted background selection, investigated in a recent study by Frisch et al. (1999b), can be considered as a special case of our treatise on combined foreground and background selection by setting $\delta_1 = \delta_2 = 0$. In this case, the target gene cosegregates perfectly with the marker alleles and indirect selection simplifies to direct selection. Consequently, only three of the nine genotypes listed for two foreground selection markers in Table 1 can occur; together with genotype 0, they correspond exactly to Types 1 to 4 defined by Frisch et al. (1999b). Moreover, because $p_{u,T+}$ is identical with $p_{u+,T+}$, the ordering of the elements in G_0 based on Eq. [18], reduce to the ordering proposed for a two-generation, marker-assisted backcross program for a dominant target gene (Frisch et al., 1999b). Furthermore, any probability of success q_2 can be reached with each of these four genotypes with $i_g = 1$ and a suitable family size f_g calculated according to Eq. [15]. These family sizes correspond to the numbers $n_1^{(2)} \ldots n_4^{(2)}$ for a two-generation background selection program given by Frisch et al. (1999b).

Rationale of the Breeding Plan

Besides the selection strategy, the core of the proposed breeding plan is (i) the definition of the subset G of marker genotypes considered as promising parents for producing generation BC_2 and (ii) the definition of the subset T of marker genotypes, which satisfy necessary conditions for a successful outcome of the backcross program. Marker genotypes with $M_l + M_r = 0$ are not included in G because with high probability they do not carry the target gene. Furthermore, homozygosity at both foreground selection markers for the recurrent parent allele results in $p_{\omega,T+} = 0$ for all $\omega \in \Omega_A$ with $M_l + M_r = 0$ and, hence, $q_\omega(i_\omega, f_\omega) = 0$ for arbitrary i_ω and f_ω. Likewise, genotypes with $Y_l + Y_r = 2$ are excluded from G because with respect to the goal of reducing the donor genome around the target gene, they show no improvement compared with F_1 individuals. In addition, they may have lost the target gene. In Appendix B, we give a mathematical proof that for all $\omega \in \Omega_A$ with $Y_l = Y_r = 2$ and arbitrary i and f the probability $q_\omega(i, f) < q_0(i, f)$ (i.e., each of these genotypes performs worse than F_1 individuals in producing BC progeny with genotype $t^+ \in T^+$).

The definition of G is also closely related to the question of how to proceed if no individual with genotype $g \in G$ is found in generation BC_1. In principle, one can either back up one generation and use an F_1 individual or backcross a BC individual with $Y_l + Y_r = 2$. Two aspects must be considered in this choice: (i) F_1 individuals carry with probability 1 the target gene, while for BC_1 individuals with $\delta_1 > 0$ and $\delta_2 > 0$ the probability $p_{\omega+|0,\omega} < 1$ for $\omega \in \Omega_A$ and $Y_l + Y_r = 2$ and (ii) F_1 individuals have on the noncarrier chromosomes an expected proportion of the recurrent parent genome of 0.50 compared to 0.75 for BC_1 individuals. Hence, with two tightly linked foreground selection markers and a BC individual with genotype $y_l^+ m_l^+ m_r^+ y_r^+$, the advantage of a higher recurrent parent genome proportion on the noncarrier chromosomes may be worthwhile to be taken at the cost of the small risk of loosing the target gene. However, when only genotypes $y_l^+ m_l^- m_r^+ y_r^+$ or $y_l^+ m_l^+ m_r^- y_r^+$ are found in BC_1, backing up one generation may be more appropriate. In this treatise, we concentrate on the region around the target gene and defined $G_0 = G \cup \{0\}$. However, replacing this definition by $G_0 = G \cup \{y_l^+ m_l^+ m_r^+ y_r^+\}$ permits using the given framework of equations for breeding programs, in which BC_1 individuals of genotype $y_l^+ m_l^+ m_r^+ y_r^+$ are preferred over F_1 individuals.

Individuals with $Y_l + Y_r = 0$ and $M_l + M_r \geq 1$ form the set T. Homozygosity at both background selection markers warrants a donor chromosome segment smaller than $d_1 + d_2$. While this applies also to the marker genotypes $y_l^+ m_l^- m_r^+ y_r^-$ and $y_l^- m_l^+ m_r^- y_r^+$, they are excluded from T because heterozygosity at a background selection marker indicates a second donor chromosome segment tightly linked to the target gene and, hence, the ultimate goal of reducing the donor genome around the target gene is not achieved by these genotypes.

Selection Strategy

The ranking of genotypes $g \in G$ according to Eq. [18] warrants a maximum q_g for $i_g = 1$ and $f_g = 1$ because

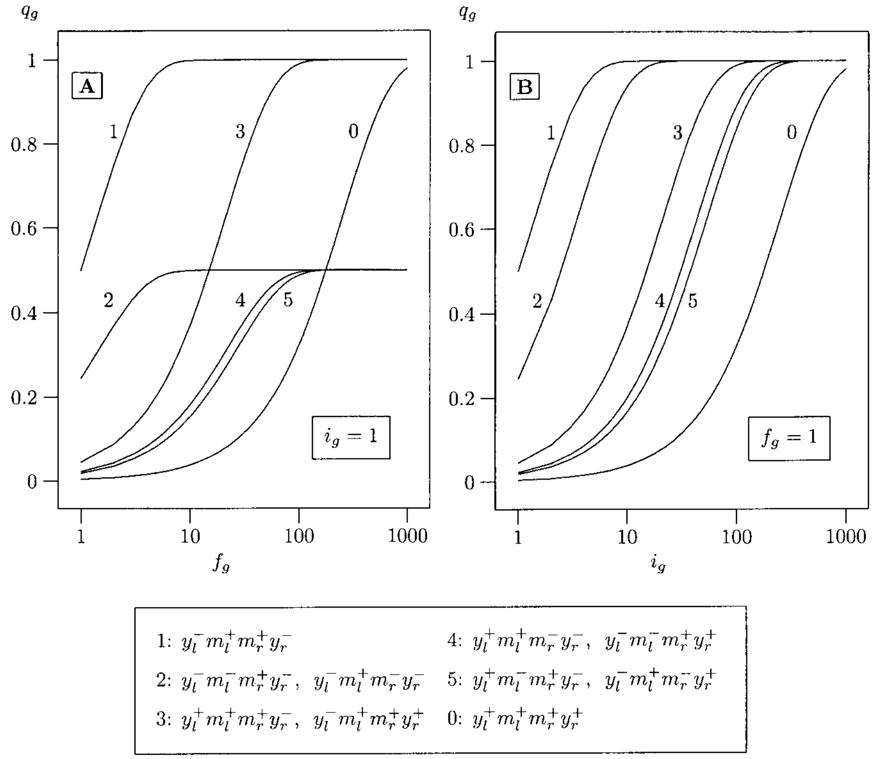

Fig. 2. Probability q_g that at least one BC progeny with genotype $t^+ \in T^+$ is generated by backcrossing individuals of Genotype $g \in G_0$. In the left diagram (A), the family size f_g derived from one backcrossed individual ($i_g = 1$) is increased. In the right diagram (B), the number of backcrossed individuals i_g is increased for family size $f_g = 1$. Marker distances are $d_1 = d_2 = 0.10$ M and $\delta_1 = \delta_2 = 0.02$ M.

in this case, Eq. [15] reduces to $p_{g+|0,g}p_{g,T+}$ which equals $p_{g,T+}$. However, for $i_u > 1$ and/or $f_u > 1$, the ranking of the genotypes with respect to the value of q reached with a certain $n_{2g} = i_g f_g$ does not necessarily remain constant. For a given i_g and large family sizes f_g, the probability q_g converges to $1 - B(i_u, 0, p_{u+|0,u})$, while for a fixed f_g and increasing i_g the probability q_g converges to 1.

This is illustrated in Fig. 2 for marker distances $d_1 = d_2 = 0.10$ M and $\delta_1 = \delta_2 = 0.02$ M. For $i_g = 1$ and increasing f_g, the probability q_g converges to 0.50 and 0.99 for BC$_1$ individuals with $M_l + M_r = 1$ and $M_l + M_r = 2$, respectively, and to 1.00 for F$_1$ individuals. Up to a family size of about 10, the initial ranking of the genotypes warrants a maximum q_g, while with larger family sizes the genotypes with $M_l + M_r = 2$ and $Y_l + Y_r = 1$ reach higher q_g than genotypes with $M_l + M_r = 1$ and $Y_l + Y_r = 2$. With family sizes larger than about 200, F$_1$ individuals reach larger values for q_0 than all genotypes with $M_l + M_r = 1$. Note that the intersection of the curves for genotypes $g,g' \in G_0$ can be obtained algebraically with the aid of Eq. [15]. For $f_g = 1$ and increasing i_g, the initial ranking of the genotypes remains constant.

For selection of BC$_1$ individuals, the discussed properties of q_g have the following consequences: (i) it may not be possible to reach a desired q_2 by backcrossing only one individual, and in particular, when selecting individuals with only one heterozygous foreground selection marker, i_g must be chosen sufficiently large; (ii) for each genotype $g \in G$, there is a family size beyond which further increments in f_g result only in a marginal gain in q_g; and (iii) when choosing individuals as parents for generation BC$_2$, the comparison about which genotypes to prefer (Eq. [14] and [15]) must be made with the family sizes that will be employed in the breeding program.

Optimum Allocation of Resources

Because the number of selected individuals and the family sizes for generation BC$_2$ are determined after knowing the outcome of generation BC$_1$, these parameters can be chosen such that any desired probability of success $0 \leq q_2 < 1$ is reached, irrespective of the choice of q_1. Nevertheless, q_1 is one of the key parameters in determining the optimum design of a breeding program. Small values for q_1 result in small n_1. In consequence, the probability of finding BC$_1$ individuals which generate by further backcrossing with a high probability BC$_2$ individuals of genotype $t^+ \in T^+$ is low. Hence, in this case a large number n_2 of BC$_2$ individuals must be produced to reach a desired q_2. In contrast, large values for q_1 result in large BC$_1$ populations and consequently a high probability of finding BC$_1$ individuals which require a smaller population size n_2 to reach a certain value of q_2 in generation BC$_2$.

We investigated the effect of the choice of q_1 on the expected total number of individuals $N = n_1 + E(n_2)$ required in two-generation backcross programs with computer simulations (the computer program is provided upon request). [$E(n_2)$ is the expected population size required in generation BC$_2$.] Population sizes n_1 were chosen such that q_1 ranged from $1 - 10^{-0.5} = 0.683772$ to $1 - 10^{-5} = 0.99999$ (Eq. [12]). With a Monte-

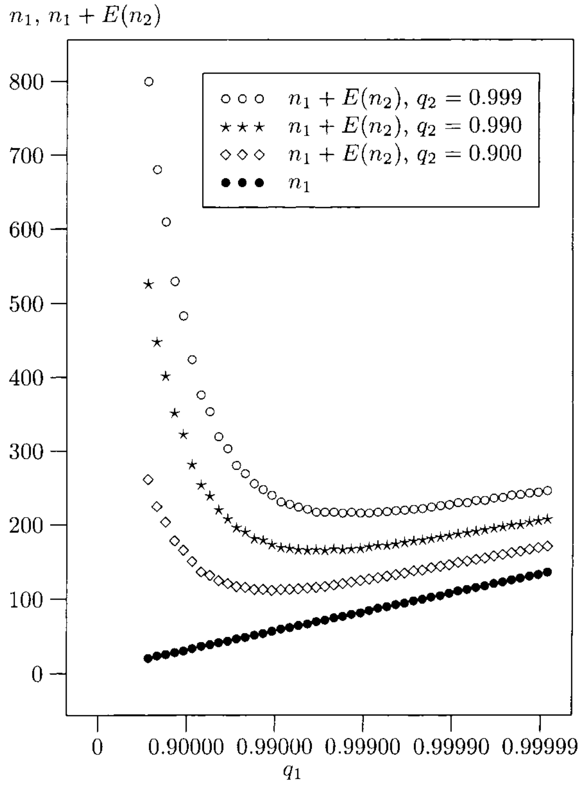

Fig. 3. Estimates of the expected total number of individuals n_1 + $E(n_2)$ required in a two generation backcross program (marker distances: $\delta_1 = \delta_2 = 0.02$ M, $d_1 = d_2 = 0.10$ M) in order to generate with probability q_2 = 0.900, 0.990, and 0.999 at least one BC$_2$ individual with genotype $t^+ \in T^+$. The values depend on the probability q_1 of obtaining at least one BC$_1$ individual with genotype $g^+ \in G^+$.

Table 2. Estimates of the optimum population size n_1^* and the respective expected population size $E(n_2)$ required to obtain with probability q_2 = 0.99 at least one BC$_2$ individual with genotype $t^+ \in T^+$ in two-generation backcross programs with one foreground and two background selection markers. The values of n_1^* and $E(n_2)$ depend on the marker distances d_1, d_2, and δ_2 (see Fig. 1).

		d_1†					
d_2	δ_2	0.04	0.06	0.08	0.10	0.15	0.20
		$n_1^*/E(n_2)$					
0.04	0.00	146/236	115/169	102/135	88/122	78/100	77/85
	0.01	208/275	165/197	142/163	124/150	113/127	111/110
0.06	0.00	122/160	93/155	76/124	68/107	61/77	51/72
	0.01	152/251	126/174	112/132	94/117	81/89	68/86
	0.03	246/280	192/209	162/179	151/154	125/138	120/125
0.08	0.00	101/135	83/118	74/109	60/95	53/68	46/59
	0.01	130/233	100/170	85/130	76/108	66/77	55/70
	0.03	169/264	140/183	119/148	110/122	91/99	80/90
	0.05	263/297	223/228	189/189	176/162	153/139	137/132
0.10	0.00	88/123	71/103	63/93	55/90	45/65	41/52
	0.01	111/209	90/161	76/124	66/103	51/78	48/62
	0.03	132/258	110/177	98/137	87/115	72/87	65/74
	0.05	178/282	143/207	130/158	120/132	105/101	88/99
0.15	0.00	80/95	56/82	49/71	43/66	38/57	31/47
	0.01	89/180	75/124	61/106	51/96	42/67	36/56
	0.03	93/242	79/174	70/138	62/105	51/77	45/63
	0.05	105/281	92/178	84/141	72/119	62.85	55/71
0.20	0.00	66/99	53/70	44/60	40/53	31/48	26/45
	0.01	71/177	62/114	53/91	47/79	36/65	30/52
	0.03	74/225	63/156	56/124	52/104	40/73	37/57
	0.05	78/256	70/178	64/142	54/111	47/79	42/64

† Marker distances d_1, d_2, and δ_2 in Morgan units.

Carlo method, we generated for each value of n_1 20 000 BC$_1$ populations. For each population the parameter space $\mathcal{A}$ was searched for the combination $(i_g, f_g)_{g \in G_0}$ requiring the minimum number of individuals n_2. The results were averaged over the repetitions in order to obtain an estimate of $E(n_2)$.

In a first series of simulations, we determined $E(n_2)$ required to reach q_2 values of 0.900, 0.990, and 0.999 for marker distances $d_1 = d_2 = 0.10$ M and $\delta_1 = \delta_2 = 0.02$ M. For $q_2 = 0.900$ and $q_2 = 0.990$, the optimum values $q_1^* = 0.987$ and $q_1^* = 0.994$ minimizing N were greater than the respective value of q_2, whereas for $q_2 = 0.999$ the optimum value $q_1^* = 0.998$ was smaller than q_2 (Fig. 3). However, for values of $q_2 = 0.990$ and $q_2 = 0.999$ the slope of the graphs were small and the choice $q_1 = q_2$ resulted in a design which required only a few more individuals than the optimum design. This shows that the obvious choice $q_1 = q_2$ has in general no optimum properties with respect to N, but was fairly close to the optimum for $q_2 = 0.990$ and $q_2 = 0.999$.

With a second series of simulations, we determined optimum values n_1^* minimizing N for varying marker distances and probability $q_2 = 0.99$. For the investigated combinations of marker distances d_1, d_2, δ_1, and δ_2, the optimum design required larger populations in generation BC$_2$ than in generation BC$_1$ irrespective of whether one or two background selection markers were employed (Tables 2 and 3). For constant d_1 and d_2, the ratio n_1:$E(n_2)$ increases with increasing δ_1 and δ_2. Tight linkage between the target gene and foreground selection markers was important with respect to the total number of individuals required. For example, a breeding program for introgression of a target gene in the center of a 20-cM background selection marker bracket required on average a total of 145 individuals when the target gene was completely linked to the foreground selection marker ($\delta_1 = 0$) (Table 2). Almost the same number of individuals (147) are required when two foreground selection markers with distance 1 cM are used (Table 3), because the probability of double crossovers in a 2-cM chromosome region is very low. However, with only one foreground selection marker located 5 cM distant from the target gene, a total of 252 individuals is required (Table 2).

A short background selection marker bracket requires considerably more individuals than a larger one. For example, with one foreground selection marker at distance $\delta_2 = 0.05$ M and $d_1 = d_2 = 0.20$ M, the optimum design requires a total of 106 individuals, whereas with $d_1 = d_2 = 0.10$ M, a total of 252 individuals are required (Table 2). This reflects that high expenditures are required for obtaining an individual with a short donor chromosome segment around the target gene.

Instead of using the total number of required individuals as criterion for optimization, one could also optimize the breeding program such that the total number of marker data points is minimal. We choose the first criterion, because when using background selection for reducing the donor chromosome segment around the target gene, the difference in the required number of marker data points for alternative parameter settings

Table 3. Estimates of the optimum population size n_1^* and the respective expected population size $E(n_2)$ required to obtain with probability q_2 = 0.99 at least one BC$_2$ individual with genotype $t^+ \in T^+$ in two-generation backcross programs with two foreground and two background selection markers. The values of n_1^* and $E(n_2)$ depend on the marker distances D_1, d_2, δ_1, and δ_2 (see Fig. 1).

		d_1†														
		0.04	0.06		0.08			0.10			0.15			0.20		
		δ_1	δ_1		δ_1			δ_1			δ_1			δ_1		
d_2	δ_2	0.01	0.01	0.03	0.01	0.03	0.05	0.01	0.03	0.05	0.01	0.03	0.05	0.01	0.03	0.05
		$n_1^*/E(n_2)$														
0.04	0.01	175/250	134/177	151/180	116/143	127/143	137/145	99/129	107/131	117/137	87/109	93/110	96/115	80/103	83/105	84/108
0.06	0.01		105/159	118/163	87/126	98/127	105/130	79/105	84/109	88/112	68/82	70/84	70/85	61/74	61/74	62/75
	0.03			140/180	100/130	116/140	129/153	94/116	102/118	113/124	69/102	75/102	80/105	62/91	74/93	70/95
0.08	0.01				75/116	78/122	83/127	65/98	71/100	74/103	51/71	58/75	60/75	49/63	49/63	51/64
	0.03					88/133	105/135	71/101	82/105	88/109	57/82	62/82	66/84	52/69	53/72	57/73
	0.05						122/156	74/108	85/117	108/118	64/85	69/90	74/97	54/77	58/88	59/89
0.10	0.01							54/93	63/95	66/97	46/67	48/69	49/73	43/54	43/56	43/58
	0.03								66/102	77/102	51/73	52/73	57/75	44/60	45/63	47/65
	0.05									84/111	55/75	59/77	62/83	46/65	49/68	53/70
0.15	0.01										38/59	40/61	42/62	33/48	35/48	35/50
	0.03											43/62	46/64	35/49	36/51	37/54
	0.05												48/69	37/52	37/55	39/58
0.20	0.01													27/44	30/45	31/45
	0.03														31/47	31/49
	0.05															32/52

† Marker distances d_1, d_2, δ_1, and δ_2 in Morgan units.

is small and DNA extraction is the major cost factor. Furthermore, because of new developments in marker technologies we expect that the cost of marker assays further reduces in the future and, hence, optimization for the required number of individuals is more important from an economical point of view.

Numerical Example

We demonstrate the application of our approach in a breeding program with a numerical step-by-step example. The first decision concerns the flanking marker distances. In general, small flanking marker distances are advantageous because (i) heterozygosity at tightly linked foreground selection markers results in a high probability that an individual carries the target gene and (ii) homozygosity at tightly linked background selection markers results in a short donor chromosome segment around the target gene. Here, we consider using the marker distances and probability of success from the example in section "Comparison with Earlier Studies" ($d_1 = d_2 = 0.1$ M, $\delta_1 = 0.03$ M, $\delta_2 = 0.05$ M, and $q_2 = 0.99$) and a maximum possible family size of $m = 200$.

We choose the population size $n_1 = 77$ because this value minimizes the expected total number $n_1 + E(n_2)$ of individuals required for the gene introgression program (Table 3). Let us assume, we marker-assayed the BC$_1$ population and found the numbers $(o_g)_{g \in G}$ of individuals with marker genotype g, which are listed in Table 4. We now rank the observed marker genotypes according to Eq. 18 (Ranking 1) and 19 (Ranking 2). For Ranking 2 we use $i_g = 1$ and $f_g = 102$, because this corresponds to the expected population size $E(n_2) = 102$ (Table 3) under the considered parameter settings. Marker genotype $y_l^- m_l^+ m_r^- y_r^-$ is most favorable under Ranking 1, while under Ranking 2 marker genotype $y_l^+ m_l^+ m_r^+ y_r^-$ is most favorable. Therefore, we first consider backcrossing individuals from these two genotypes.

We try to find the smallest family size for which $q_2 \geq 0.99$ when selecting exactly one individual of either of these marker genotypes and no individual of any other marker genotype. While for $(i_g, f_g) = (1, 10)$ the marker genotype $y_l^- m_l^+ m_r^- y_r^-$ yields a higher q_2 value than $y_l^+ m_l^+ m_r^+ y_r^-$, it is not possible to reach $q_2 \geq 0.62$ by backcrossing only one individual of marker genotype $y_l^- m_l^+ m_r^- y_r^-$, even when using the maximum family size $m = 200$ (Table 5). (Note that for this marker genotype $P_{g+|0,g} = 0.62$ (Table 4), see also the discussion of Fig.

Table 4. Parameters for generation BC$_1$ in the numerical example (for a detailed description and definitions of symbols see text). r_1 and r_2 are the ranks of the observed marker genotypes according to Eq. 18 and 19, respectively.

Marker genotype $g \in G$	$p_{0,g}$	$p_{g+\|0,g}$	$p_{g+,T+}$	o_g	r_1	r_2
$y_l^- m_l^+ m_r^+ y_r^-$†	0.0014	0.9985	0.4993	0		
$y_l^- m_l^- m_r^+ y_r^-$	0.0016	0.3751	0.4762	0		
$y_l^- m_l^+ m_r^- y_r^-$	0.0023	0.6249	0.4854	1	1	2
$y_l^+ m_l^+ m_r^+ y_r^-$	0.0206	0.9985	0.0447	2	2	1
$y_l^- m_l^+ m_r^+ y_r^+$	0.0288	0.9985	0.0446	3	3	3
$y_l^+ m_l^- m_r^+ y_r^-$	0.0001	0.3751	0.0453	0		
$y_l^+ m_l^+ m_r^- y_r^-$	0.0329	0.6249	0.0317	3	4	4
$y_l^- m_l^- m_r^+ y_r^+$	0.0329	0.3751	0.0432	2	5	5
$y_l^- m_l^+ m_r^- y_r^+$	0.0001	0.6249	0.0440	0		

† The symbols y_t and y_r denote the background selection markers, and m_o and m_r the foreground selection markers. A superscript + or − indicates that the locus is heterozygous or homozygous for the recurrent parent allele, respectively.

Table 5. Alternative selection parameters (i_g, f_g) in the numerical example (for a detailed description see text) and the resulting population sizes n_2 (Eq. [16]) and probabilities of success q (Eq. [14]).

Marker genotype g			
$y_l^- m_l^+ m_r^- y_r^-$†	$y_l^+ m_l^+ m_r^+ y_r^-$	n_2	q
(i_g, f_g)			
(1,1)		1	0.30
(1,10)		10	0.62
	(1,1)	1	0.04
	(1,10)	10	0.36
	(1,105)	105	0.99
	(2,51)	102	0.99
(1,10)	(2,40)	90	0.99
(1,9)	(1,81)	90	0.99

† The symbols y_l and y_r denote the background selection markers, and m_1 and m_r the foreground selection markers. A superscript + or − indicates that the locus is heterozygous or homozygous for the recurrent parent allele, respectively.

2). The minimum population size to reach $q_2 \geq 0.99$ by backcrossing one individual of marker genotype $y_l^+ m_l^+ m_r^+ y_r^-$ is $n_2 = 105$. This number is reduced to $n_2 = 102$ when backcrossing two instead of one individual of marker genotype $y_l^+ m_l^+ m_r^+ y_r^-$.

Now we investigate parameter combinations where individuals of two marker genotypes are selected. The previous calculations showed that for $f_g = 10$ the marker genotype $y_l^- m_l^+ m_r^- y_r^-$ results in greater q_2 values than the marker genotype $y_l^+ m_l^+ m_r^+ y_r^-$. Therefore, we choose $(n_g, f_g) = (1, 10)$ for $g = y_l^- m_l^+ m_r^- y_r^-$. In combination with $(n_g, f_g) = (2, 40)$ for $g = y_l^+ m_l^+ m_r^+ y_r^-$, a probability $q_2 \geq 0.99$ is reached with $n_2 = 90$ (Table 5). Also $(n_g, f_g) = (1, 9)$ for $g = y_l^- m_l^+ m_r^- y_r^-$, in combination with $(n_g, f_g) = (1, 81)$ for $g = y_l^+ m_l^+ m_r^+ y_r^-$ reaches $q_2 \geq 0.99$ with $n_2 = 90$. For marker genotypes $g = y_l^+ m_l^+ m_r^+ y_r^-$ and $g = y_l^- m_l^+ m_r^+ y_r^+$, the probabilities $p_{g,T+}$ are almost identical (Table 4). Hence, the parameter setting $(n_g, f_g) = (1, 9)$ for $g = y_l^- m_l^+ m_r^- y_r^-$ and $(n_g, f_g) = (1, 81)$ for $g = y_l^- m_l^+ m_r^+ y_r^+$ reaches also $q_2 \geq 0.99$ with $n_2 = 90$.

Consequently, an optimum selection strategy is to backcross the BC_1 individual with marker genotype $y_l^- m_l^+ m_r^- y_r^-$ with a family size of 9 individuals and to backcross one out of the five BC_1 individuals with marker genotypes $y_l^+ m_l^+ m_r^+ y_r^-$ *or* $y_l^- m_l^+ m_r^+ y_r^+$ with a family size of 81 individuals. These selection parameters combine a high selection intensity with a minimum number of required individuals.

REFERENCES

Allard, R.W. 1960. Principles of Plant Breeding. John Wiley and Sons, Inc., New York.

Bosch, K. 1993. Statistik-Taschenbuch. Oldenburg Verlag, Munich, Germany.

Browning, S. 2000. The relationship between count-location and stationary renewal models for the chiasma process. Genetics 155:1955–1960.

Frisch, M., M. Bohn, and A.E. Melchinger. 1999a. Comparison of selection strategies for marker-assisted backcrossing of a gene. Crop Sci. 39:1295–1301.

Frisch, M., M. Bohn, and A.E. Melchinger. 1999b. Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. Crop Sci. 39:967–975. Erratum in Crop Sci. 39:1903.

Haldane, J.B.S. 1919. The combination of linkage values and the calculation of distances between linked factors. J. Genet. 8:299–309.

Hospital, F., and A. Charcosset. 1997. Marker assisted introgression of quantitative trait loci. Genetics 147:1469–1485.

Hospital, F., C. Chevalet, and P. Mulsant. 1992. Using markers in gene introgression breeding programs. Genetics 132:1199–1210.

Karlin, S., and U. Liberman. 1978. Classification of multilocus recombination distributions. Proc. Nat. Acad. Sci. (USA) 75:6332–6336.

Kosambi, D.D. 1944. The estimation of the map distance from recombination values. Ann. Eugen. 12:172–175.

McPeek, M.S., and T.P. Speed. 1995. Modeling interference in genetic recombination. Genetics 139:1031–1044.

Melchinger, A.E. 1990. Use of molecular markers in breeding for oligogenic disease resistance. Plant Breed. 104:1–19.

Morgan, T.H., C.B. Bridges, and J. Schulz. 1935. Constitution of the germinal material in relation to heredity. Carnegie Inst. Washington Publ. 34:282–291.

Openshaw, S.J., S.G. Jarboe, and W.D. Beavis. 1994. Marker-assisted selection in backcross breeding. p. 41–43. *In* Proceedings of the Symposium "Analysis of Molecular Marker Data," Corvallis, OR. 5–6 Aug. 1994. Am. Soc. Hortic. Sci., Alexandria, VA, and CSSA, Madison, WI.

Stam, P. 1979. Interference in genetic crossing over and chromosome mapping. Genetics 92:873–594.

Stam, P., and A.C. Zeven. 1981. The theoretical proportion of the donor genome is near-isogenic lines of self-fertilizers bred by backcrossing. Euphytica 30:227–238.

Tanksley, S.D. 1983. Molecular markers in plant breeding. Plant Mol. Biol. Rep. 1:1–3.

Visscher, P.M., C.S. Haley, and R. Thompson. 1996. Marker-assisted introgression in backcross breeding programs. Genetics 144:1923–1932.

Young, N.D., and S.D. Tanksley. 1989. RFLP analysis of the size of chromosomal segments retained around the *tm*-2 locus of tomato during backcross breeding. Theor. Appl. Genet. 77:353–259.

Zhao, H., and T.P. Speed. 1996. On genetic map functions. Genetics 142:1369–1377.

APPENDIX A

Obtaining *k* Individuals of Genotype $g^+ \in G^+$ or $t^+ \in T^+$

The probability g of finding in a sample of n at least k individuals of a genotype, which occurs with probability p, is

$$q = 1 \sum_{m=0}^{k-1} B(n, m, p)$$

$$= \mathrm{F}_{2k,2(n-k+1)} \left(\frac{p(n-k+1)}{(1-p)k} \right) \qquad [20]$$

(Bosch, 1993, p. 296), where F is the cumulative density function of the F-distribution. By defining c as the q percentile of the F distribution with parameters $2k$ and $2(n-k+1)$,

$$c = \mathrm{F}^{-1}_{2k,2(n-k+1)}(q) = \frac{p(n-k+1)}{(1-p)k}, \qquad [21]$$

we obtain the minimum population size n required to find with probability q at least k individuals as

$$n \geq c \frac{1-p}{p} k + k - 1. \qquad [22]$$

This result can be used to generalize the presented approach in order to generate in generation BC_1 at least k individuals of genotype $g^+ \in G^+$ and/or in generation BC_2 at least k individual of genotype $t^+ \in T^+$. In generation BC_1, the minimum sample size to generate with probability q_1 at least k individuals of genotype $g^+ \in G^+$ is obtained by inserting in Eq. [21] and [22]:

$$p = \sum_{g^+ \in G^+} p_{0,g+}$$

$$q = q_1. \qquad [23]$$

For a given $(i_g, f_g)_{g \in G_0}$, the probability of recovering at least k BC_2 individuals of genotypes $t^+ \in T^+$ is

$$q[(i_g, f_g)_{g \in G_0}] = 1 - \sum_{z=0}^{k-1} P(Z = z), \qquad [24]$$

where

$$P(Z = z) = \sum_{z_0 + \cdots + z_g = k} \prod_{g \in G} P(Z_g = z_g) \qquad [25]$$

and

$$P(Z_g = z_g) = \sum_{s=1}^{i_g} [B(i_g, s, p_{g+|0,g}) \times B(sf_g, x_g, p_{g+,T+})]. \qquad [26]$$

Equation [24] can be used instead of Eq. [14] to compare alternative parameter settings.

APPENDIX B

Proof of the proposition: For each genotype $\omega \in \Omega_A$ with $Y_1 + Y_2 = 2$, the probability $q_\omega(i, f) < q_0(i, f)$.

Case 1, $\omega = y_l^+ m_l^- m_r^- y_r^+$:

From $p_{\omega+,T+} = 0$ follows $q_\omega(i, f) = 0$. q.e.d.

For the remaining two cases, we make use of the fact that $B(i_\omega, s, p_{\omega+|0,\omega})$ is an increasing function of $p_{\omega+|0,\omega}$ and $1 - B(sf_\omega, 0, p_{\omega+,T+})$ is an increasing function of $p_{\omega+,T+}$. Hence, Eq. [15] implies that the proposition holds true if $p_{\omega+|0,\omega} < 1$ and $p_{\omega+,T+} \leq p_{0,T+}$.

Case 2, $\omega = y_l^+ m_l^- m_r^+ y_r^+$: We have

$$p_{\omega+|0,\omega} = \frac{p_c(1 - p_d)}{p_h} = \frac{p_c(1 - p_d)}{p_c + p_d - 2p_c p_d}$$
$$= \frac{p_c(1 - p_d)}{p_c(1 - p_d) + (1 - p_c)p_d} < 1$$
$$\text{for } \delta_1, \delta_2 > 0$$

and

$$p_{\omega+,T+} = p_{\omega+,t+} \qquad \text{with } t^+ = y_l^- m_l^- x^+ m_r^+ y_r^-$$
$$= p_a(1 - p_d)p_e/2 = p_b(1 - p_c)(1 - p_d)p_e/2$$
$$+ (1 - p_b)p_c(1 - p_d)p_e/2 \leq p_{0,T+},$$

using Eq. [9]. q.e.d.

For symmetry reasons the proposition holds also true for $\omega = y_l^+ m_l^+ m_r^- y_r^+$.

Case 3, $\omega = y_l^+ m_l^+ m_r^+ y_r^+$: We have

$$p_{\omega+|0,\omega} = \frac{(1 - p_c)(1 - p_d)}{(1 - p_h)}$$
$$= \frac{(1 - p_c)(1 - p_d)}{(1 - p_c)(1 - p_d) + p_c p_d} < 1$$
$$\text{for } \delta_1, \delta_2 < 0$$

and $p_{\omega+,T+} = p_{0,T+}$. q.e.d.

Detection and Mapping of a Major Locus for Fusarium Wilt Resistance in Common Bean

A. L. Fall, P. F. Byrne,* G. Jung, D. P. Coyne, M. A. Brick, and H. F. Schwartz

ABSTRACT

Fusarium oxysporum Schlectend. Fr. f. sp. _phaseoli_ J.B. Kendrick and W.C. Snyder (FOP) is a vascular pathogen that causes Fusarium wilt in common bean (_Phaseolus vulgaris_ L.). This disease is an increasing problem in the western U.S., and exploitation of genetic resistance is considered the most feasible control method. The objective of this study was to detect quantitative trait loci (QTL) for Fusarium wilt resistance in a population derived from an interracial cross between FOP-susceptible Belneb RR-1 (race Durango) × FOP-resistant A55 (race Mesoamerica). Seventy-six F_6-derived recombinant inbred lines (RILs) were screened for disease severity in greenhouse inoculations and rated on a scale of 1 (resistant) to 9 (susceptible). The phenotypic data were compared to existing random amplified polymorphic DNA (RAPD) marker data using single-factor analysis of variance. Marker U20.750 on linkage group (LG) 10 accounted for 63.5% of the phenotypic variance for this trait. Lines exhibiting the A55 banding pattern at this locus had disease severity scores that averaged 3.6 points lower than lines with the Belneb RR-1 pattern. Two additional markers, AD4.450 on LG 3 and K10.700 on LG 11, were significant (_P_ < 0.01) in single-factor analysis of variance, but only marker U20.750 on LG 10 remained significant when composite interval mapping (CIM) was conducted. The tight linkage between the putative QTL and U20.750, as indicated by CIM, makes this marker a promising candidate for conversion to a sequence-characterized amplified region (SCAR) for use in marker-assisted selection in Fusarium wilt resistant common bean cultivar development.

FUSARIUM WILT is a vascular disease of common bean. This fungus has been detected in bean-growing regions throughout the world, and is an economically significant problem in Latin America, Africa, and the western United States (Pastor-Corrales and Abawi, 1987; Buruchara and Camacho, 2000; Salgado et al., 1995). Infected plants display symptoms of yellowing, wilting, and necrosis of leaf and stem tissue, which often results in hastened maturity, decreased seed size, and yield loss (Schwartz et al., 1996). In the High Plains region of the U.S., yield losses in fields affected with Fusarium wilt are estimated at 10 to 30% (Salgado et al., 1995). Fusarium wilt is difficult to control due to formation of chlamydospores that remain viable in soil for long periods of time. Chemical, cultural, and biocontrol treatments cannot effectively limit this disease, making genetic resistance the most viable control measure.

Genetic resistance to FOP race 4 has been identified in germplasm from common bean races Durango and Mesoamerica of the Middle American gene pool (Velasquez-Valle et al., 1997). In some populations of Durango, inheritance of resistance to this isolate is controlled by a single dominant gene, designated *Fop4* (Salgado et al., 1995; Cross et al., 2000). However, attempts to identify a resistance locus using bulked segregant analysis have been unsuccessful in two Durango populations (Cross, 1998). In Mesoamerica, there are indications of polygenic resistance to Fusarium wilt (Salgado et al., 1995; Cross et al., 2000).

A.L. Fall, P.F. Byrne, and M.A. Brick, Dep. of Soil and Crop Sciences, Colorado State Univ., Ft. Collins, CO 80523; G. Jung, Dep. of Plant Pathology, Univ. of Wisconsin, Madison, WI 53706; D.P. Coyne, Dep. of Horticulture, Univ. of Nebraska, Lincoln, NE 68583; H.F. Schwartz, Dep. of Bioagricultural Sciences and Pest Management, Colorado State Univ., Ft. Collins, CO 80523. Published as Paper no. 13212 Journal Series, Nebraska Agricultural Division, Univ. of Nebraska, Lincoln, NE. Received 9 Nov. 2000. *Corresponding author (pbyrne@lamar.colostate.edu).

Published in Crop Sci. 41:1494–1498 (2001).

Abbreviations: CIM, composite interval mapping; DSI, disease severity index; FOP, *Fusarium oxysporum* f. sp. *phaseoli*; LG, linkage group; LOD, log of the odds; RAPD, random amplified polymorphic DNA; RIL, recombinant inbred line; SCAR, sequence-characterized amplified region; QTL, quantitative trait locus (loci).